VG	(non infinite) verb group	[VG was_VBDZ beaten_VVN VG]
com	comment phrase	<com Well_UH com>
vpp	verb with prepositional particle	[vpp of_IO |_| [VG handling_VVG VG] vpp]
pp	prepositional phrase	[pp in_II (NR practice_NN1 NR) pp]
NR	noun phrase (non referent)	(NR Dinamo_NP1 Kiev_NP1 NR)
R	noun phrase (referent)	(R it_PPH1 R)
WH	wh-word phrase	(WH which_DDQ WH)
QNT	quantifier phrase	<QNT much_DA1 QNT>
ADV	adverb phrase	<ADV still_RR ADV>
WHADV	wh-adverb phrase	<WHADV when_RRQ WHADV>
ADJ	adjective phrase	<ADJ prone_JJ ADJ>

Table 2

The process then divides the input sentence into elements. Chunks are regarded as elements, as are sentence markers, paragraph markers, punctuation marks and words which do not fall inside chunks. Each chunk has a marker applied to it which identifies it as a chunk. These markers constitute chunk markers 99.

The output from the chunking process for the above example sentence is shown in Table 3 below, each line of that table representing an element, and 'phrasetag' representing a chunk marker.

SENTSTART
phrasetag(<ADV) Similarly_RR
'_'
phrasetag((NR) Britain_NP1
phrasetag([VG) became_VVD
phrasetag(<ADJ) popular_JJ
phrasetag[pp after_ICS (NR a_AT1 rumour_NN1 NR) pp]
phrasetag[VG got_VVD about_RP VG]

that_CST
phrasetag(NR Mrs_NNSB1 Thatcher_NP1 NR)
phrasetag[VG had_VHD declared_VVN VG]
phrasetag(NR open_JJ house_NNL1 NR)
SENTEND
._.

Table 3

The computer then carries out classification process 100 under control of the program. The classification process 100 uses a classification of words and pronunciation database 100A. The classification database 100A is the fifth of the five databases stored on the CD-ROM 32.

The classification database is divided into classes which broadly correspond to parts-of-speech. For example, verbs, adverbs and adjectives are classes of words. Punctuation is also treated as a class of words. The classification is hierarchical, so many of the classes of words are themselves divided into sub-classes. The sub-classes contain a number of word categories which correspond to the word tags 95 applied to words in the input text 40 by the parsing process 94. Some of the sub-classes contain only one member, so they are not divided further. Part of the classification (the part relating to verbs, prepositions and punctuation) used in the present embodiment is given in Table 4 below.

verbs	&FW	
	BTO22	
	EX	
	II22	
	RA	
	RGR	
	beverbs	VB0 VBDR VBG VBM VBN VBR VBZ

	doverbs	VD0 VDG VDN VDZ
	haveverbs	VH0 VHG VHN VHZ
	auxiliary	VM VM22 VMK
	baseform	VV0
	presentpart	VVG
	past	VBDZ VDD VHD VVD VVN
	thirdsingular	VVZ
	verbpart	RP
prepositions	iopp	IO
	iwpp	IW
	icspp	ICS
	iipp	II
	ifpp	IF
punctuation	minpunct	comma rhtbrk leftbrk quote ellipsis dash
	majpunct	period colon exclam semicol quest

Table 4

It will be seen that the left-hand column of Table 4 contains the classes, the central column contains the sub-classes and the right-hand column contains the word categories. Figure 4 shows part of the classification of verbs. The class of words 'verbs' includes four sub-classes, one of which contains only the word category 'RP'. The other sub-classes ('beverbs', 'doverbs', and 'past') each contain a plurality of word categories. For example, the sub-class 'doverbs' contains the word categories corresponding to the word tags VDO, VDG, VDN, and VDZ.

In carrying out the classification process 100 the computer first identifies a core word contained within each chunk in the input text 40. The core word in a prepositional chunk (i.e. one labelled 'pp' or 'vpp') is the first preposition within the chunk. The core word in a chunk labelled 'WH' or 'WHADV' is the first word in the chunk. In all other types of chunk, the core word is the last word in the chunk. The